Modelo para la detección de cáncer de seno en imágenes histológicas a partir de aprendizaje profundo con múltiples anotadores

J. M. Buitrago Chavez[1] J. F. González Velez[2] J. Gil González[3]

5 de agosto de 2024

[1]*Pontificia Universidad Javeriana Cali, Colombia*
[2]*Pontificia Universidad Javeriana Cali, Colombia*
[3]*Pontificia Universidad Javeriana Cali, Colombia*

Resumen

El cáncer de seno es el tipo de cáncer más común en mujeres, y su detección temprana requiere comúnmente de la ayuda de un experto para su correcto diagnóstico. Por lo anterior, el aprendizaje profundo puede ser una herramienta de gran ayuda para este problema; sin embargo, es difícil encontrar un conjunto de datos robusto y bien etiquetado que permita un buen entrenamiento de los modelos. En este proyecto se proponen diferentes modelos de aprendizaje profundo para la clasificación de imágenes histológicas, aprovechando las anotaciones de múltiples personas con diversos grados de experiencia y conocimiento. Los resultados de estos modelos fueron comparados con varios modelos clásicos de aprendizaje automático que utilizan las etiquetas verdaderas para su entrenamiento.

1. Introducción

El cáncer, especialmente el cáncer de seno, es una de las mayores causas de muerte a nivel mundial. La detección temprana de esta enfermedad es crucial para mejorar las posibilidades de recuperación de los pacientes afectados, sin embargo, las pruebas diagnósticas tradicionales pueden no ser suficientemente sensibles. Es por esto que la tecnología, mediante el uso de métodos de aprendizaje profundo, emerge como una herramienta prometedora para la detección temprana del cáncer de seno.

El aprendizaje profundo es una solución prometedora para detectar cáncer de seno en imágenes histológicas, automatizando y mejorando la precisión del análisis. Sin embargo, se enfrenta al desafío de la escasez de datos de alta calidad debido a la necesidad de expertos para anotar las imágenes, lo que limita la cantidad de información relevante disponible.

En este proyecto, los datos provienen de un proceso de "crowdsourcing" estructurado con colaboradores de diversos niveles de conocimiento en patología para anotar imágenes histológicas del TCGA (The Cancer Genome Atlas) [1]. Se seleccionaron 151 regiones de interés con la ayuda de patólogos senior, resultando en 75,243 imágenes tras el preprocesamiento para el entrenamiento de los modelos.

El proyecto propone diferentes modelos de aprendizaje automático basados en anotaciones de múltiples personas, comparándolos con un modelo tradicional que utiliza etiquetas verdaderas. Estos modelos emplean "transfer learning" con la red VGG16 para mejorar la precisión y eficiencia del análisis de imágenes histológicas de cáncer de seno.

1.1. Objetivos

El objetivo principal de este proyecto fue construir un modelo de aprendizaje profundo basado en múltiples anotadores para la detección de cáncer de seno en imágenes histológicas. Para esto, se tuvieron en cuenta los siguientes objetivos específicos:

- Implementar una etapa de preparación del conjunto de datos a partir de técnicas de procesamiento de datos, con el fin de que pueda ser utilizado adecuadamente por los modelos de aprendizaje profundo para la detección de cáncer de seno en imágenes histológicas.
- Implementar un modelo de aprendizaje profundo que tenga en cuenta múltiples anotadores para la clasificación de imágenes histológicas.
- Implementar un proceso de búsqueda de hiperparámetros para el modelo desarrollado.
- Validar los resultados del modelo de aprendizaje profundo con múltiples anotadores.

2. Fundamentación Teórica

Los problemas de aprendizaje supervisado de imágenes con un único anotador consisten en un

conjunto de entrenamiento $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, donde $\mathbf{x}_i \in \mathbb{N}^{M \times P \times C}$ es una matriz de $M \times P$ píxeles y C canales de la imagen. En problemas de clasificación multiclase $y_i \in \{1, ..., K\}$ es la etiqueta que indica la clase correspondiente.

La etiqueta y_i, conocida como "ground truth" o "gold standard", debe ser proporcionada por un experto en el área. En medicina y análisis de imágenes histológicas, este proceso es laborioso y consume mucho tiempo para un solo anotador [2].

El *crowdsourcing* permite obtener múltiples etiquetas de distintos anotadores con una mayor agilidad, pero introduce problemas debido a las diferencias en conocimiento y sesgos de los anotadores, lo que añade ruido al conjunto de datos. Además, puede haber anotadores maliciosos que etiqueten incorrectamente a propósito. Es crucial considerar estos problemas al entrenar modelos utilizando múltiples anotadores, especialmente ajustando las funciones de costo en redes neuronales profundas, de tal forma que puedan manejar el ruido y sesgos de múltiples personas. [2, 3].

La función de costo *Generalized Categorical Cross-entropy* (GCCE) se presenta como una posible solución al poder entrenar sobre múltiples anotadores y ajustar la resistencia al ruido. Esta función además asigna un parámetro $\lambda^r \in [0, 1]$ a cada anotador r, que representa la confiabilidad o peso de la etiqueta de dicho anotador. Estos parámetros λ se entrenan en conjunto con el resto del modelo, lo que permite ajustar el peso de las opiniones de cada anotador durante el entrenamiento. Además, la función tiene el hiperparámetro $q \in (0, 1]$, el cual a mayor valor, mayor resistencia al ruido presenta la función. [4]

Una solución similar a GCCE, es *Crowdlayer*, la cuál es una capa propuesta en el artículo "Deep Learning from Crowds" que busca aprender directamente de las etiquetas de múltiples anotadores mediante diversas estrategias para calcular sesgos y pesos de cada anotador por clase. Esta capa se agrega al final de un modelo antes del entrenamiento. Utiliza entropía cruzada como función de costo, ajustada para manejar etiquetas faltantes asignando un valor de 0. Tras el entrenamiento, se elimina la capa *Crowdlayer*, dejando la capa de salida expuesta para evaluar el rendimiento del modelo. *Crowdlayer* maneja sesgos y pesos de los anotadores mediante cinco configuraciones: matriz de pesos (MW), vector de pesos por clase (VW), vector de sesgos por clase (VB), combinación de VW y VB (VW+B), y un único peso por anotador (SW). [3]

Focal loss es otra función de costo con propiedades interesantes. Esta es una modificación de *cross-entropy* que le asigna menor penalización a las muestras bien clasificadas por el modelo, lo cual permite darle mayor importancia a las muestras mal clasificadas. Además, la función puede darle un mayor peso a una muestra dependiendo de la clase verdadera a la que pertenezca, lo que permite asignar diferentes prioridades a cada clase del conjunto de datos. [5]

La función de *focal loss* no puede utilizar múltiples anotadores, por lo que es necesario reducir múltiples etiquetas a una única etiqueta. La votación mayoritaria o *majority voting*, en su forma más simple, consiste en utilizar la moda de las etiquetas de una muestra para determinar una única etiqueta. Existen variaciones de este método que le asignan diferentes pesos a cada anotador dependiendo de su experiencia o nivel de conocimiento.

3. Resultados

3.1. Obtención y preprocesamiento de las imágenes histológicas

Los datos de este proyecto se obtuvieron del estudio "Structured crowdsourcing enables convolutional segmentation of histology images"[6]. La investigación implementa un método de crowdsourcing estructurado, involucrando a participantes con diversos niveles de conocimiento para anotar imágenes histológicas de cáncer de seno obtenidas del Cancer Genome Atlas (TCGA)[1]. Se seleccionaron 151 imágenes histológicas de casos confirmados de cáncer de seno, donde los patólogos senior identificaron las regiones de interés. A su vez, los estudiantes de medicina, en conjunto con los patólogos junior, realizaron las anotaciones, con revisiones y correcciones a cargo de los patólogos senior. Las imágenes anotadas se almacenaron como máscaras en escala de grises, clasificadas en categorías como tumor, estroma e infiltración linfocítica.

Después del proceso descrito anteriormente, las imágenes resultantes variaban en su tamaño y espacio de color, por lo que es necesario preprocesar las imágenes de la siguiente manera:

- **Normalización de color:** Mediante el método de normalización de color de Reinhard[7] se alteran los diferentes canales de colores de las imágenes según la media y la desviación estándar de los canales de color de una imagen de muestra. Obteniendo un dataset de imágenes coherente para el entrenamiento de los diferentes modelos.
- **Segmentación de las imágenes:** Dado que las imágenes del conjunto de datos varían en tamaño, se implementa un proceso de segmentación utilizando una ventana deslizante de 224x224 píxeles. Este método extrae segmentos que contienen al menos un 50 % de píxeles

de alguna clase específica, asegurando uniformidad en las dimensiones y aumentando la cantidad de muestras útiles para el entrenamiento de los modelos.

3.2. Extracción de características

Se utilizó una VGG16 preentrenada para extraer características de las imágenes histológicas. Este proceso, realizado una sola vez, reduce significativamente el tiempo y los recursos computacionales necesarios para el entrenamiento de modelos. La estructura de la CNN incluye la VGG16 sin la capa "fully connected", seguida de una capa "average pooling" y una capa "flatten", transformando cada imagen en un vector de características para su uso en el entrenamiento.

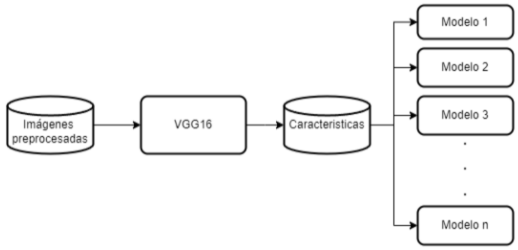

Figura 1: Flujo de extracción y utilización de características.

3.3. Gold standard

Estos modelos fueron entrenados con la etiqueta verdadera y sirven como una base comparativa para medir la eficacia de los modelos entrenados con múltiples anotadores. Se utilizaron 2 funciones de costo: *categorical cross-entropy* y *focal loss*. *Focal loss* tiene 2 hiperparámetros, γ y α. Para γ se realizó una búsqueda de hiperparámetros de 9 valores: 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5 y 5.0. Los valores de α se utilizaron para asignar mayor peso a aquellas muestras con menor representación en el conjunto de datos, donde $\alpha = [0.1678, 0.2259, 0.6061]$. Además, se entrenaron modelos sin usar α para comprobar el impacto del hiperparámetro.

Accuracy	F1 class 1	F1 class 2	F1 class 3
0.8165	0.8679	0.7382	0.7359

Cuadro 1: Resultado del modelo gold standard con la función de costo cross-entropy

γ	Accuracy	F1 class 1	F1 class 2	F1 class 3
1.0	0.8155	0.8687	0.7462	0.7148
1.5	0.7942	0.8501	0.7229	0.6905
2.0	0.7886	0.8478	0.7179	0.6738
2.5	0.8106	0.8640	0.7372	0.7176
3.0	0.7949	0.8505	0.7215	0.6973
3.5	0.7982	0.8571	0.7214	0.6900
4.0	0.7615	0.8127	0.6921	0.6763
4.5	0.7842	0.8409	0.7087	0.6852
5.0	0.7924	0.8553	0.6998	0.6840

Cuadro 2: Resultados del modelo gold standard con la función de costo focal loss usando el parámetro α.

γ	Accuracy	F1 class 1	F1 class 2	F1 class 3
1.0	0.8250	0.8753	0.7434	0.7451
1.5	0.8154	0.8661	0.7380	0.7369
2.0	0.8262	0.8766	0.7457	0.7383
2.5	0.8213	0.8705	0.7412	0.7481
3.0	0.8138	0.8682	0.7295	0.7079
3.5	0.8199	0.8766	0.7260	0.7058
4.0	0.8121	0.8683	0.7305	0.6923
4.5	0.8175	0.8704	0.7329	0.7248
5.0	0.8039	0.8586	0.7139	0.7178

Cuadro 3: Resultado del modelo gold standard con la función de costo focal loss sin usar el parámetro α

3.4. Majority Voting

Estos modelos utilizan *majority voting* para obtener una única etiqueta. Se utilizaron las mismas funciones de costo e hiperparámetros utilizados en *gold standard*.

Accuracy	F1 class 1	F1 class 2	F1 class 3
0.8206	0.8706	0.7425	0.7422

Cuadro 4: Resultados del modelo con *majority voting* usando la función de costo *cross-entropy*

γ	Accuracy	F1 class 1	F1 class 2	F1 class 3
1.0	0.7818	0.8412	0.7082	0.6797
1.5	0.7715	0.8340	0.6891	0.6758
2.0	0.7982	0.8578	0.7110	0.7039
2.5	0.7649	0.8216	0.6976	0.6723
3.0	0.7843	0.8474	0.6973	0.6836
3.5	0.7682	0.8258	0.6799	0.6844
4.0	0.7538	0.8168	0.6478	0.6420
4.5	0.7686	0.8271	0.6878	0.6889
5.0	0.7469	0.7991	0.6559	0.6997

Cuadro 5: Resultado de modelos con *majority voting* con la función de costo *focal-loss* usando el parámetro α

γ	Accuracy	F1 class 1	F1 class 2	F1 class 3
1.0	0.8190	0.8720	0.7350	0.7242
1.5	0.8060	0.8607	0.7221	0.7115
2.0	0.7968	0.8491	0.7142	0.7093
2.5	0.8103	0.8697	0.7093	0.7017
3.0	0.7996	0.8575	0.7077	0.6901
3.5	0.8113	0.8678	0.7199	0.7070
4.0	0.7961	0.8559	0.6977	0.6866
4.5	0.8000	0.8594	0.7047	0.6874
5.0	0.7889	0.8540	0.6899	0.6611

Cuadro 6: Resultado de modelos con *majority voting* con la función de costo *focal-loss* sin usar el parámetro α

3.5. Generalized Categorical Cross-entropy

Se utilizó la función de costo GCCE para entrenar múltiples modelos. Esta función de costo tiene un hiperparámetro q, del cual se realizó una búsqueda entre 10 valores: 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Se entrenaron 2 versiones, una utiliza todo el conjunto de datos, mientras que la otra utiliza un conjunto reducido (2140 muestras), donde hay al menos 2 etiquetas por anotador.

q	Accuracy	F1 class 1	F1 class 2	F1 class 3
0.01	0.7610	0.7814	0.7028	0.6496
0.1	0.8148	0.8654	0.7372	0.7289
0.2	0.8123	0.8634	0.7355	0.7271
0.3	0.8076	0.8593	0.7350	0.7058
0.4	0.8100	0.8629	0.7385	0.7043
0.5	0.7623	0.8294	0.6706	0.5230
0.6	0.7027	0.7452	0.6315	0.3473
0.7	0.6734	0.6689	0.6347	0.4325
0.8	0.7392	0.7979	0.6881	0.4032
0.9	0.7436	0.7733	0.6652	0.5058

Cuadro 7: Resultados de modelos con GCCE usando todo el conjunto de datos.

q	Accuracy	F1 class 1	F1 class 2	F1 class 3
0.01	0.7926	0.8464	0.7287	0.6221
0.1	0.8050	0.8608	0.7357	0.6109
0.2	0.8004	0.8563	0.7270	0.6260
0.3	0.7980	0.8523	0.7285	0.6295
0.4	0.7972	0.8515	0.7300	0.6184
0.5	0.8001	0.8557	0.7240	0.6218
0.6	0.7978	0.8515	0.7319	0.6188
0.7	0.8014	0.8595	0.7377	0.5463
0.8	0.8007	0.8565	0.7365	0.5837
0.9	0.7994	0.8580	0.7330	0.5514

Cuadro 8: Resultados de modelos con GCCE usando los datos que fueron anotados por dos o más personas.

3.6. Crowdlayer

La capa *crowdlayer* cuenta con 5 estrategias para manejar el sesgo y/o pesos de los anotadores. Se entrenaron modelos con cada una de las 5 estrategias para comparar sus rendimientos. Además, de manera similar a GCCE, se entrenaron 2 versiones, una con el conjunto de datos completo y otra con el conjunto de datos reducido de 2140 muestras.

Conn type	Accuracy	F1 class 1	F1 class 2	F1 class 3
MW	0.7514	0.8282	0.6754	0.4990
SW	0.8184	0.8671	0.7430	0.7428
VB	0.8176	0.8662	0.7420	0.7533
VW	0.7654	0.8087	0.7005	0.7342
VW+B	0.8022	0.8523	0.7343	0.7215

Cuadro 9: Resultados de modelos con *Crowdlayer* usando todo el conjunto de datos.

Conn type	Accuracy	F1 class 1	F1 class 2	F1 class 3
MW	0.7965	0.8498	0.7299	0.6259
SW	0.8023	0.8563	0.7339	0.6311
VB	0.8051	0.8595	0.7337	0.6372
VW	0.7964	0.8511	0.7305	0.6192
VW+B	0.7968	0.8525	0.7228	0.6223

Cuadro 10: Resultados de modelos con *Crowdlayer* usando los datos que fueron anotados por dos o más personas.

3.7. Análisis

Entre todos los modelos evaluados en el desarrollo de este proyecto, el modelo de *majority voting* utilizando la función de costo *cross-entropy* (tabla 4) muestra el mejor rendimiento global con una exactitud de 0.8206 ± 0.0106 y puntajes F1 equilibrados. Aunque algunos modelos específicos con diferentes configuraciones de GCCE y Crowdlayer se aproximan a estos resultados, ninguno logra superarlo.

Sin embargo, el *majority voting* puede no ser el método más confiable para manejar múltiples anotadores, puesto que asume que todos los anotadores son igualmente confiables y no poseen sesgos significativos. Las estrategias de *generalized categorical cross-entropy* y *crowdlayer* solucionan esta dificultad al calcular los pesos y sesgos de cada anotador durante el entrenamiento.

Es importante tomar en cuenta las características del conjunto de datos cuando se trabaja con múltiples anotadores. Si bien es posible que **majority voting** proporcione resultados competentes, es importante explorar otros métodos que tomen en cuenta la posibilidad de presencia de ruido o sesgos en las anotaciones.

4. Discusión y Conclusiones

Al tratarse de un problema de clasificación de imágenes es necesario contar con un conjunto de datos robusto que permita la correcta extracción de características y por ende un entrenamiento óptimo de los diferentes modelos estudiados a lo largo de este proyecto. por ende, estudios como "Structured crowdsourcing enables convolutional segmentation of histology images"[6] y el artículo "Deep Gaussian Processes for Classification With Multiple Noisy Annotators. Application to Breast Cancer Tissue Classification"[8] presentaron una base sólida y un punto de partida para dar solución al problema propuesto.

Para conseguir la correcta clasificación de las imágenes por parte de los modelos propuestos fue necesario manipular internamente los mismos, diseñando e implementando funciones de costo personalizadas para abordar diferentes desafíos como:

- **Manejo de muestras difíciles de clasificar:** La función de costo *Focal Loss* se diseñó para abordar el problema de las muestras difíciles de clasificar y el desequilibrio de clases. Esta función de costo Asigna mayor peso a las muestras mal clasificadas, mejorando la discriminación entre clases.
- **Manejo de múltiples anotadores:** La función de costo *Generalized Categorical Cross Entropy* permite el aprendizaje a partir de anotaciones de múltiples individuos con diferentes niveles de conocimiento. Incorpora un parámetro de ponderación que ajusta la contribución de cada anotador, optimizando el aprendizaje de anotaciones confiables y minimizando el impacto de las erróneas.

En conclusión, el proyecto se centró en desarrollar modelos de aprendizaje automático para detectar cáncer de seno en imágenes histológicas, aprovechando las opiniones de múltiples individuos con distintos niveles de conocimiento. Estos modelos se compararon con un modelo ideal entrenado con etiquetas verdaderas, demostrando que es posible lograr resultados similares. Los resultados mostraron que los modelos que integraron anotaciones de diversos individuos alcanzaron un rendimiento comparable al del modelo ideal, e incluso superaron ligeramente este punto de referencia en algunos casos, indicando que la diversidad en las anotaciones puede mejorar la capacidad del modelo para identificar el cáncer de seno.

Referencias

[1] N. C. Institute, "Using tcga data," https://www.cancer.gov/ccg/research/genome-sequencing/tcga/using-tcga-data, Accessed 2023.

[2] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, no. 43, pp. 1297–1322, 2010. [Online]. Available: http://jmlr.org/papers/v11/raykar10a.html

[3] F. Rodrigues and F. Pereira, "Deep learning from crowds," 2017.

[4] J. C. Triana-Martinez, J. Gil-González, J. A. Fernandez-Gallego, A. M. Álvarez Meza, and C. G. Castellanos-Dominguez, "Chained deep learning using generalized cross-entropy for multiple annotators classification," *Sensors*, vol. 23, no. 7, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/7/3518

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2018.

[6] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. T. Elsebaie, L. S. Abo Elnasr, R. A. Sakr, H. S. E. Salem, A. F. Ismail, A. M. Saad, J. Ahmed, M. A. T. Elsebaie, M. Rahman, I. A. Ruhban, N. M. Elgazar, Y. Alagha, M. H. Osman, A. M. Alhusseiny, M. M. Khalaf, A.-A. F. Younes, A. Abdulkarim, D. M. Younes, A. M. Gadallah, A. M. Elkashash, S. Y. Fala, B. M. Zaki, J. Beezley, D. R. Chittajallu, D. Manthey, D. A. Gutman, and L. A. D. Cooper, "Structured crowdsourcing enables convolutional segmentation of histology images," *Bioinformatics*, vol. 35, no. 18, pp. 3461–3467, 02 2019. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz083

[7] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, 2001.

[8] M. López-Pérez, P. Morales-Álvarez, L. A. D. Cooper, R. Molina, and A. K. Katsaggelos, "Deep gaussian processes for classification with multiple noisy annotators. application to breast cancer tissue classification," *IEEE Access*, vol. 11, pp. 6922–6934, 2023.